Customer Churn Analysis

Problem Statement:

Domain Topic Telecom Churn Analysis Telecom Churn (loss of customers to competition) is a problem for telecom companies because it is expensive to acquire a new customer and companies want to retain their existing customers. Most telecom companies suffer from voluntary churn.

Customer churn is when a company's customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.

Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can priorities focused marketing efforts on that subset of their customer base.

Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

I will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

Evaluation Data Analysis

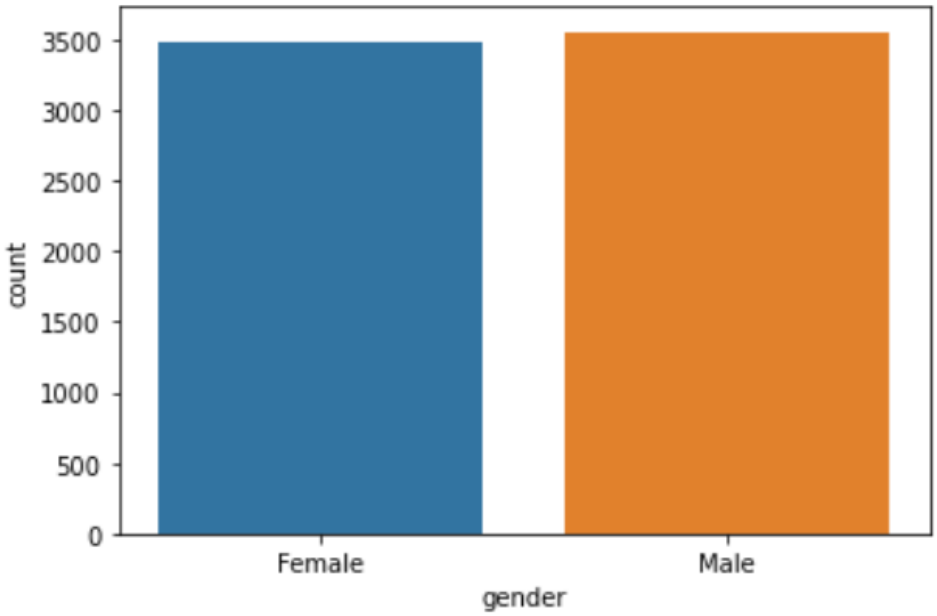
Data comprises of Services — which services the customer subscribed to (internet, phone, cable, etc.), Monthly bill total, Tenure — How long they had been a customer, Basic demographic info — whether they were elderly, had dependents, etc.

let us do the data analysis one by one

Gender

As the gender indicate , that how much mostly , in telecom industry the customer has the insured sex count and from my analysis

```
sns.countplot(data['gender'])
plt.show()
```




```
data['gender'].value_counts()
```

```
Male      3555
Female    3488
Name: gender, dtype: int64
```

it clearly shows that telecom industry does not depend upon gender and has almost equal ratio between male and female, only uniqueness in dataset

SENIOR-CITIZEN

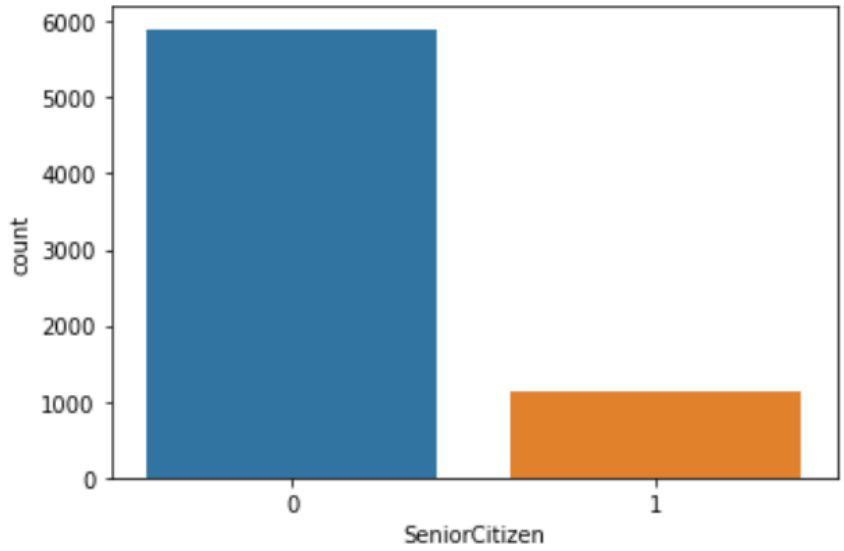
it can be a possible exploratory analysis that senior citizen , if they are in count, can possible can leave doing business with the company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer but

as a senior citizen cannot that be possibly true for keeping customer for a long time

```
In [25]: data['SeniorCitizen'].value_counts()

Out[25]: 0    5901
         1    1142
         Name: SeniorCitizen, dtype: int64

In [26]: sns.countplot(data['SeniorCitizen'])
         plt.show()
```



maximum people are not senior citizen and only 16.2% people from data are senior citizen but still 16% have possibility of leaving company or not using much of services.

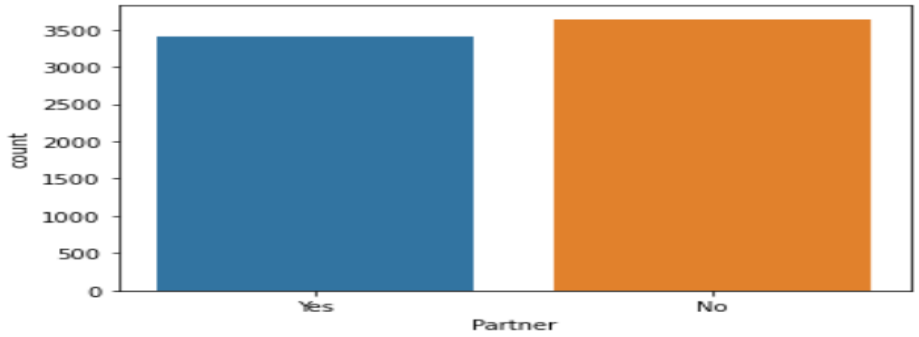
PARTNER

Partner Interconnect provides connectivity between your on-premises network and your Virtual Private Cloud (VPC) network through a supported service provider.

```
data['Partner'].value_counts()
```

```
No      3641
Yes     3402
Name: Partner, dtype: int64
```

```
sns.countplot(data['Partner'])
plt.show()
```



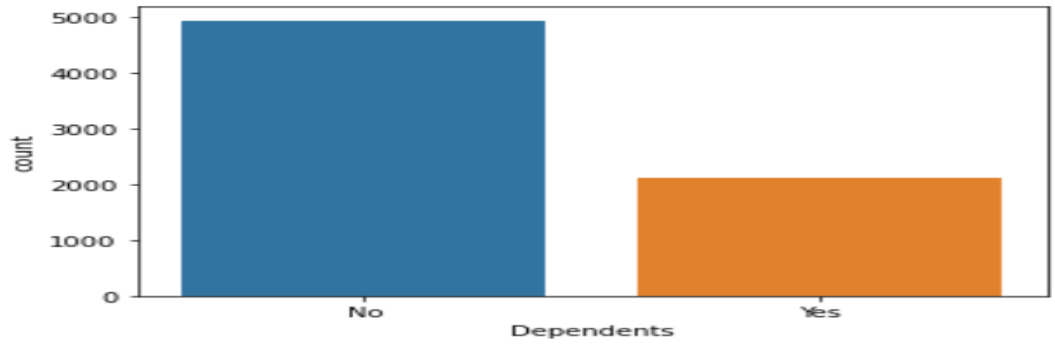
DEPENDENTS

It surely means whether the customer is likely to be dependent or not, **Customer** without **dependents** are four times more **likely to churn**. Senior citizens are three times less likely to churn

```
data['Dependents'].value_counts()
```

```
No      4933
Yes     2110
Name: Dependents, dtype: int64
```

```
sns.countplot(data['Dependents'])
plt.show()
```



Most of the customers are independent. and almost 30 percent are dependent upon others, as possible explanation can be , that maximum people earn by themselves and pay bills, can only count students and senior citizen in this 30 percent

TENURE

Tenure refers to the number of months that a **customer** has subscribed for. The **tenure** for a churning **customer** indicates the

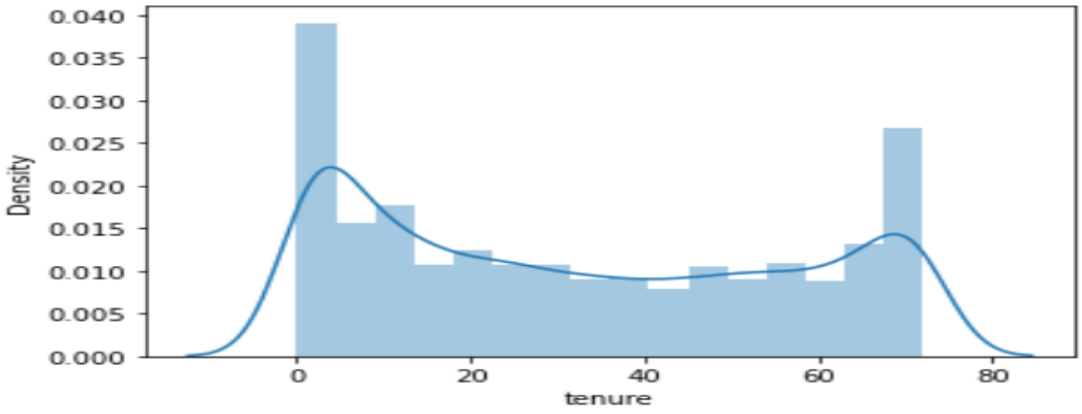
number of months that they spent as a subscriber.so on an average, tenure of a customer can be taken as 32 which is good , so even after this much tenure , if a customer is leaving, may be there can be lot of other services that encouraging customer to leave.

TENURE/CHURN (BI-vairiate analysis)

From he analysis, the customers with less tenures mostly stop the services.

The regular customers are distributed uniformly, have normal distribution

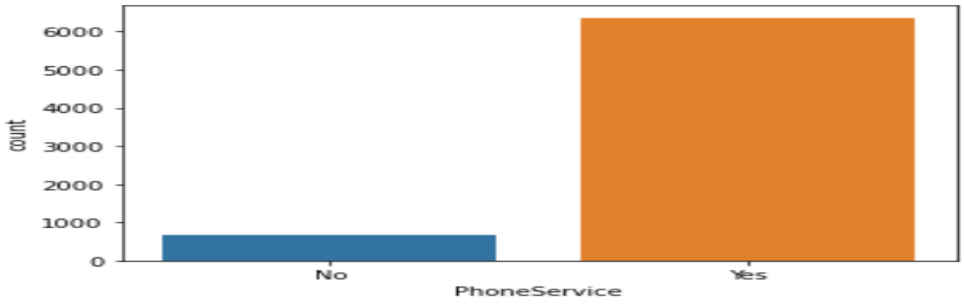
```
sns.distplot(data['tenure'])
plt.show()
```



PHONE SERVICE

A **phone service** is a public utility **service** focused on voice communications delivered by a **phone** company to residential and commercial clients.

```
[34]: sns.countplot(data['PhoneService'])
      plt.show()
```



maximum people have almost 90 % have phone service and 10 percent doesn't have this service. Out of 7000, 700 don't have phone services, might be using other services only for once or twice.

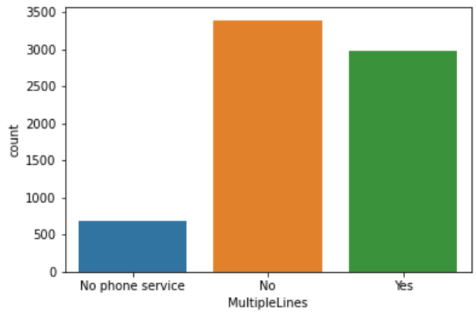
MULTIPLE LINES

A **Multiline** is a group of analogue telephone **lines** with one telephone number. The first **line** is known as the main exchange **line**, whilst the extra **lines** are known as secondary **lines**. The additional **lines** are added to the main exchange telephone **line** so there's no need for a new number.

```
data['MultipleLines'].value_counts()
```

```
No                  3390
Yes                 2971
No phone service     682
Name: MultipleLines, dtype: int64
```

```
sns.countplot(data['MultipleLines'])
plt.show()
```



other than the people with no phone service at all, Customers equally use the multiline and single line phone services.

INTERNET SERVICES

Telecom service providers offer voice calling and text messaging directly, while **internet services** run on the top layer of the network as standalone **services**. The **services** are offered in a different context, even if they sometimes serve the same function for end

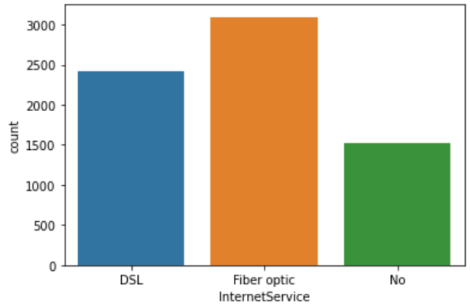
Most of the customers have opted for Fibre Optic as it is the fasted and highest bandwidth connumication meduim.

Fibre optic being able to have a very high speed, the cost for the fibre optics is high, thus many of the custmers also prefer DSL i.e., Digital Subscriber Line.

Yet 21 percent of the customers prefer not have any internet connection.

```
print(data['InternetService'].value_counts())
sns.countplot(data['InternetService'])
plt.show()
```

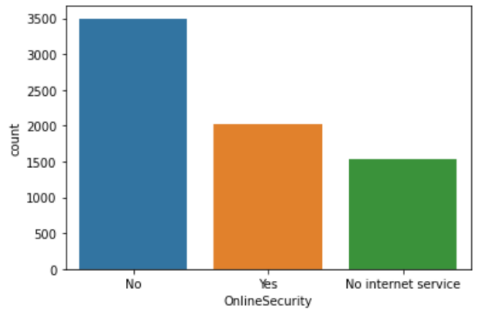
```
Fiber optic    3096
DSL            2421
No             1526
Name: InternetService, dtype: int64
```



ONLINE SECURITY

Online security refers to the body of technologies, processes, and practices designed to protect networks, devices, programs, and data from attack, damage, or unauthorized access. **Online security** may also be referred to as information technology **security**.

```
sns.countplot(data['OnlineSecurity'])
plt.show()
```




```
data['OnlineSecurity'].value_counts()
```

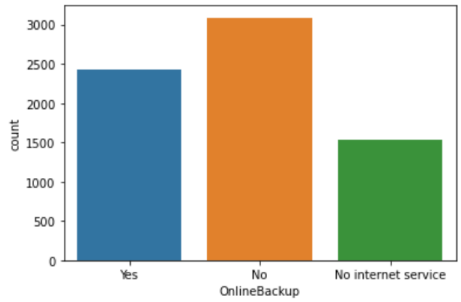
```
No                     3498
Yes                    2019
No internet service    1526
Name: OnlineSecurity, dtype: int64
```

only 21% people have online security, others dont have.

ONLINE BACKUP

A remote **backup**, **online backup** or **cloud backup** is where a copy of computer files or **systems** are stored in a secure offsite location. In the event of an IT failure where the original data is lost, the remotely stored data can be used to return operations to normal running.

```
sns.countplot(data['OnlineBackup'])
plt.show()
```




```
data['OnlineBackup'].value_counts()
```

```
No                     3088
Yes                    2429
No internet service    1526
Name: OnlineBackup, dtype: int64
```

only 34% have online backup options with them, rest don't have it.

'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport','StreamingTV', 'StreamingMovies'

There above are the services on can avail only if they had availed the internet service.

the above services have almost same percentage as above two shown, shows that if the customer has internet services available, the only he can avail other services like 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport','StreamingTV', 'StreamingMovies'

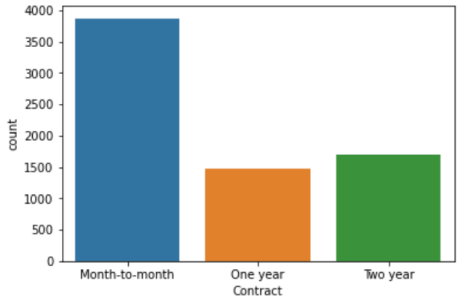
ALL SERVICES ABOVE/MONTHLY CHARGES

From the analysis, its shown that the more services customer has, the more will be monthly be added, that is **The monthly charges increase as the number of sub products increase**

CONTRACT

If you are a **telecom** business or if you are contracting for **telecommunication** services, your **contract** is likely to be highly technical and possibly very **detailed**.

```
sns.countplot(data['Contract'])
plt.show()
```




```
data['Contract'].value_counts()
```

```
Month-to-month    3875
Two year          1695
One year          1473
Name: Contract, dtype: int64
```

```
#3875 customer are not have month-to-month contract and 1695 have two year contract and 1473 have one-year contract.
```

As you can see from above count, Customer mostly prefer the month to month contract, possible options as for customer it is difficult to pay all amount all of a sudden for year or two.

PAYMENT METHODS

A customer can make **payment** using different **payment methods** that are supported by the service provider; for example, the customer can make **payments** using the **payment methods** such as cheque, credit card, debit card or wire transfers, or direct cash deposit

from analysis, electronic cheque is preferred by customers as they feel safe to pay for the services

PAYMENT METHOD VS TENURE

AS people with high tenure with company , customers do prefer automatic payments more.

MONTHLY CHARGES AND TOTAL CHARGES

as per analysis goes by, monthly charges for a customer from box plot , shows between 40–80/- and total charges aand maximum between 0–2000/- and then goes down with count.

INTERNET SERVICES /MONTHLY CHARGES

AS most people prefer month to month charges but monthly wise Fibre optics are costly than DSL

MULTIPLE LINES/MONTHLY CHARGES

AS most people prefer month to month charges but multiples lines make monthly charge costly

TOTAL CHARGES/CHURN

from the analysis, total charges/monthly charges are almost vice versa and uniformly distributed to other features

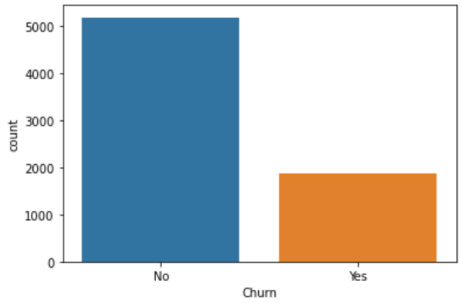
CHURN

Churn rate is the percentage of subscribers to a service that discontinue their subscription to that service in a given time period. ... **Churn** rate is an important consideration in the telephone and cell phone services industry.

```
data['Churn'].value_counts()
```

```
No     5174
Yes    1869
Name: Churn, dtype: int64
```

```
sns.countplot(data['Churn'])
plt.show()
```




```
# Not Churn Customer is more & Churn Customer is less
# There is data imbalance problem present in dataset
```

PRE PROCESSING

FROM CORRELATION

I can conclude that ***To avoid unstable estimates of coeffiecients in the model, need to drop the 'Total Charges' variable during regression process, as it is highly correlated to both 'Tenure' and 'Monthly Charges'.***

It can be seen from correlation plot that, Contract_Month-to-month, Online security_No, Tech Support_No...etc. are positively correlated with Churn. While, on the other end of the plot, tenure, Contract_Two year, InternetService_No...etc. are negatively correlated with Churn.

Interestingly, services such as Online security, Streaming TV, OnlineBackup, TechSupport..., etc. with InternetService_No seem to be negatively related to Churn. Need to explore the patterns more for the above correlations below before modelling and identifying the important variables.

NULL VALUES AND REDUNDANT COLUMNS

so over all we have almost 7043records and 21 columns

no null values and got rid of redundant columns as customer ID.

OUTLIERS

From box plot, there were no outliers as well.

SKEW STATS

As data was positively skewed, so square root tranformation worked in a positive way for balancing.

LABEL ENCODING

as maximum data was categorical so from Label encoder, dataset been encoded and for object data types as data was mostly categorical so we converted into numerical datypes

```
#  Convert object datatype into numerical dtypes
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
for j in object_dtypes:
    data[j] = le.fit_transform(data[j])
```

```
# Checking Datypes
data.dtypes
```

```
customerID            int32
gender                int32
SeniorCitizen         int64
Partner               int32
Dependents            int32
tenure                int64
PhoneService          int32
MultipleLines         int32
InternetService       int32
OnlineSecurity        int32
OnlineBackup          int32
DeviceProtection      int32
TechSupport           int32
StreamingTV           int32
StreamingMovies       int32
Contract              int32
PaperlessBilling      int32
PaymentMethod         int32
MonthlyCharges      float64
TotalCharges          int32
Churn                 int32
```

DATA IMBALANCE:

```
In [88]: # Now deal with data imbalance problem
         from imblearn.over_sampling import SMOTE
         sm=SMOTE()
         x_over,y_over = sm.fit_resample(x,y)
```

```
In [89]: y_over.value_counts()

Out[89]: 1    5174
         0    5174
         Name: Churn, dtype: int64
```

SCALING

As dataset need to get scaled to get maximum accuracy, standard scaler was taken help to get out data set properly scaled before building model

PRINCIPAL COMPONENTS ANALYSIS

These use for dimension reduction technique

MODEL BUILDING

For train_test_split we 70% data for training and remaining 30% for testing purpose

```
log = LogisticRegression()
svc = SVC()
knn = KNeighborsClassifier(n_neighbors=5)
abc=AdaBoostClassifier()
decision_tree = DecisionTreeClassifier()
rfc = RandomForestClassifier()
gnb=GaussianNB()
```

CONCLUSION

```
models = pd.DataFrame({'Classifier':['LogisticRegression', 'SVC', 'KNeighborsClassifier','AdaBoostClassifier','DecisionTreeClassi
                       'Score':[log_accuracy,svc_accuracy,knn_accuracy,abc_accuracy,dt_accuracy,rfc_accuracy,gnb_accuracy],
                        'CVS':[log_cvs,svc_cvs,knn_cvs,abc_cvs,dt_cvs,rfc_cvs,gnb_cvs],
                        'Difference':[log_Difference,svc_Difference,knn_Difference,abc_Difference,dt_Difference,rfc_Difference,gnb

})
models.sort_values(by='Score',ascending=False)
```

	Classifier	Score	CVS	Difference
5	RandomForestClassifier	0.852818	0.794691	5.812705
1	SVC	0.848309	0.734630	11.367902
3	AdaBoostClassifier	0.824799	0.800370	2.442893
6	GaussianNB	0.822544	0.755076	6.746865
0	LogisticRegression	0.819324	0.795257	2.406622
2	KNeighborsClassifier	0.797424	0.683655	11.376821
4	DecisionTreeClassifier	0.759098	0.726679	3.241919

we see that AdaBoostClassifier is best model and also give highest accuracy 82% and least difference between accuracy and cross validation

HYPERPARAMETER TUNNING:

FINAL CONCLUSION & CLASSSFICATION REPORT:

```
print(classification_report(y_test,y_predict))
```

```
              precision    recall  f1-score   support

           0       0.85      0.81      0.83      1562
           1       0.82      0.85      0.83      1543

    accuracy                           0.83      3105
   macro avg       0.83      0.83      0.83      3105
weighted avg       0.83      0.83      0.83      3105
```

```
#After hyperparatmeter tunning of of AdaBoostClassifier is 83%
```

CONCLUDING REMARKS

in above model it is summarised that relation analysis between features and target with best visualisation

also best model has been predicted with best hyperparameter tuning and best score was derived and made 83% accuracy

from the over conclusions made →

people having very high tenure or very less tenure are leaving company

people don’t have the phone services aren’t enjoying other services , so probably customer is leaving , here company can work upon new schemes so that customer can get attract towards services.

there is no issue with monthly billing or method of payment , but maximum customers cant afford for two year subscription , company should come up with again new schemes and offers.

after model building, Logistic Regression algorithm looks best for the telecom customer churn dataset, which will predict the churn analysis.